https://doi.org/10.2352/EI.2025.37.11.HVEI-209

How does the perceived quality of compressed images depend on image content?

Claire Mantel, Hui Wang, Thierry Soreze and Søren Forchhammer; DTU Electro - Department of Electrical and Photonics Engineering, Technical University of Denmark, 343 Ørsteds Plads, 2800 Kgs. Lyngby, Denmark

Abstract

The impact of image compression algorithms varies significantly across image contents in a way that is challenging to predict. The ongoing trend towards richer visual content, e.g. High Dynamic Range and Wide Color Gamut, increases both the relevance and complexity of this issue. This study analyzes first grades of perceived quality of compressed images to determine in which proportion their variance is due to compression levels, image content and compression type, respectively. An ANOVA analysis on 3 HDR datasets indicates that the variance of the subjective evaluations is due for 45-62% to the compression level and for 7-10% to the image content. Secondly, we present a framework for identifying which features calculated on the source images are efficient to predict the part of image content in grades of perceived quality of compressed images. We build on traditional regression analysis by adding an adaptation of the recent Model Class Reliance approach. In an experiment on 6 published datasets of subjective quality grades of compressed images, OLS-R and KNN models predicting the grades are built using two input variables: the compression level and one feature characterizing the original content. The Empirical Model Reliance is then calculated to measure the importance of the content feature in the regression model as well as the Model Class Reliance to bound the impact of a reduced fit to the training data, i.e. indicating robustness towards generalization. Results show that traditional regression analysis alone is not robust for identifying the most relevant features and confirms that when the most useful features for SDR are SI/block contrast measures, other features characterize HDR content best, such as DR and color features (colorfulness or saturation).

Introduction

The visual quality resulting from the compression of an image depends on the chosen compression algorithm and settings but also on the characteristics of the original content. Subjective evaluations of quality show that the content impacts significantly the perceived quality. This is illustrated on results from Korshunov et al. on HDR image compression [1]: in Figure 1 the average perceived quality as a function of the compression level differs significantly across content. On the compression side, the impact of the image content needs to be controlled to ensure that the test and evaluation of performance are not confined to the specific data they are tested on but will generalize to other content. It is also relevant to predict and select the compression settings that will yield the desired performance for each content specifically. To mitigate that uncertainty, one approach is to ensure using sufficiently varied content for performance evaluation. For video compression two indicators characterizing the source are standardized in ITU-T Rec. P.910 [2]: the Spatial perceptual information (SI) and the temporal perceptual information (TI). For image compression, it is common good practice to use Spatial perceptual information and colorfulness as two indicators representing the variations across images of a dataset [3]. For objective quality metrics too, it has been shown that one of the biggest challenges is predicting quality across different contents [4]. As both objective and subjective evaluations of perceived quality of compressed images embed the effect of compression, original content and their interaction in a single grade, isolating their respective impact is not trivial and aligning quality across content is an active research topic [5]. Another challenge comes from the trend towards richer signal content such as High Dynamic Range (HDR) or Wide Color Gamut (WCG), which in practice extends the range of possible content [5, 6] .

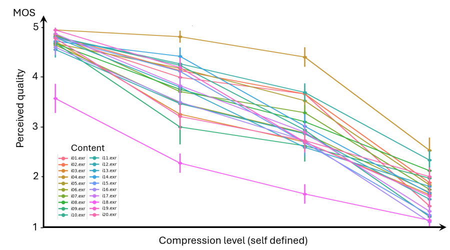


Figure 1: Perceived quality (mean opinion score - MOS) as a function of the compression level (x-axis) from [1]. Each line color corresponds to a different content. The MOS is averaged over the three compression settings.

This study focuses first on the respective impact of content and compression levels on the perceived subjective quality. An estimate of these proportions is presented and compared across publicly available datasets of subjective image quality. Secondly, we analyze further the part of variance of perceived quality due to content by relating it to specific aspects of image characteristics. The approach used here is regression analysis to evaluate the importance of features for well performing models [7]. The research questions addressed in this study are: **RQ1** What are typical proportions of the MOS due to content vs due to compression level respectively in existing QA datasets ? **RQ2** To what extent can we explain/predict the variance due to content by features representing some specific characteristics of images?

The remainder of this paper is organized as follows: the related literature is presented in the next section (State of the Art), then the proposed method is detailed in Sec. Method for analysis before being applied to 6 publicly available datasets with subjective evaluations of compressed images in Sec. Experiment.

State of the Art

Evaluation of variation within a dataset

The two indicators standardized by the ITU to characterize sources for video compression SI and TI [2] have been updated in 2022 to address the different range and coding scheme of HDR content. Although no standard currently recommends equivalent indicators for image compression, it is good practice to indicate SI and colorfulness for 8bits RGB images (measured typically by [8]) and additionally Dynamic Range and Image Key for HDR content [9]. An analysis of characteristics of publicly available quality assessment datasets was performed in [3] in terms of source content, test conditions and obtained MOS scores. The author computes SI, colorfulness (and motion vectors for videos) and calculates their respective range and uniformity (calculated as the entropy on 10 bins) to characterize the variety among source contents. This study from 2012 is limited to SDR/BT.709 content and studies separately the content and the MOS but does not investigate their interactions. A framework was designed in [10] by Narwaria et al. where the authors apply iteratively an objective measure to evaluate the impact of contrast reduction on HDR content with BT.709 primaries. The sensitivity of the content to contrast reduction is used to assess how complex it will be for tone mapping algorithms. Specifically for image compression, different studies have researched how to measure and predict how "easy" an image will be to compress. A measure of this coding efficiency (also termed compressibility or complexity) is presented in [11] as the AUC (Area Under Curve) under a MOS-rate curve. Yu et al. define the coding complexity as the number of bits needed when compressing at fixed QP [12], seen as the best approximation for Kolmogorov complexity. Determining the relevant characteristics of the perceived quality of images can also be used to focus modeling efforts. For example to predict the perception of dynamic range of HDR content (BT.709 primaries) in [9] or which display characteristics are relevant for quality on HDR displays in [13].

Evaluation of feature importance for regression models

In subjective experiments on the quality of compressed images, the measured dependent variable is the perceived quality and the independent variables are the compression levels, the content and possibly additional variables such as compression settings or repetition order. The proportion of the variance of the subjective grades due to the independent variable "content" and to the independent variable "compression level" is the effect size of the factor as calculated by N-way ANOVA [14]. However, "content" is then a categorical variable and this gives no information over which aspects of the content are key. Regression analysis can tackle that question by comparing which features input to the regression models are most useful for prediction. Common approaches are hierarchical or iterative regression in which features are added one by one to regression models (usually simple linear models) where the gain from adding a feature represents its impact. In [15], Krasula et al. use a mixture of random and sequential approach (Las Vegas algorithm) for adding features to linear regression model predicting quality of tone mapped images. Step-wise regression is applied by Hulusic et al. [9] to predict perceived dynamic range of HDR images.

Regression analysis has recently been the focus of renewed research interest in relation to explainability in AI/ML. In this study, we follow the approach formalized by Fisher et al. in [7] to evaluate the importance of a set of features of interest $\{X_i\}$ for the prediction of a dependent variable Y. Fisher et al. define a set of "well performing models" $\mathscr{F}_R = \{f_j \setminus Loss(f_j) \geq f_{ref} - \varepsilon\}$ named Rashomon set, of the regression models which performance falls within a margin ε of a "model of interest" f_{ref} that serves as reference. The authors then measure how much each model of the Rashomon set rely on the feature of interest. For a given feature of interest X_i and a fixed model f_j from the Rashomon set $\mathscr{F}_R$ this *empirical model reliance* $\widehat{MR}(X_i, f_j)$ is calculated as:

$$\widehat{MR}(X_i, f_j) = \frac{\hat{e}_{switchX_i}(f_j)}{\hat{e}_{origin}(f_j)}$$

where $\hat{e}_{origin}(f_j)$ is the expected loss for f_j and $\hat{e}_{switchX_i}(f_j)$ is the expected loss for f_j when noise is introduced on X_i. The noise introduced on X_i should remove the correspondence with the measured variable but retain the overall distribution of X_i. To this aim, the authors use permutations of X_i so $\widehat{MR}(X_i, f_j)$ measures the impact on the model performance when breaking the relation of the feature of interest X_i to the target variable Y but without modifying the distribution of X_i. Finally an *empirical model class reliance* $\widehat{MCR}$ is calculated for each feature of interest X_i as the bounds for empirical model reliance:

$$\widehat{MCR} = [\min_{f_j \in \mathscr{F}_R} \widehat{MR}(X_i, f_j),\ \max_{f_j \in \mathscr{F}_R} \widehat{MR}(X_i, f_j)]$$

It corresponds to how much the $\widehat{MR}$ can vary when relaxing the model fit but retaining a minimum performance constraint. The authors also detail how to render the calculations tractable for specific classes of regression models such as linear models used here.

Method for analysis

The complete pipeline for the study is depicted in Fig. 2. The N-way ANOVA analyzes how the independent variables, Compression level (CpLvl), Content (Cnt) and Compression type (CpTyp) impact the dependent variable, the subjective quality scores. In an ANOVA the independent variables are categorical. This section presents our method of representing the Content variable via a continuous variable, a feature characterizing the image content. The three main steps are described in the following sections: chosen features, regression analysis and estimation of the importance of feature via Model Class Reliance.

Chosen features

In a first step, various features are calculated to characterize different aspects of the source content. **Group 1 - Spatial content descriptors on luminance plane**: Spatial information [2] and contrast measures calculated on blocks: Weber contrast (CbW), Michelson contrast (CbM), RMS contrast (CbRMS) [16, 15] are used to characterize the spatial variation of content in the luma channel. **Group 2 - Amplitude related descriptors**: Dynamic range, Image Key and Area are used to characterize the amplitude of the pixel intensities for HDR content [9]. **Group 3 - Color**: 3.a Colorfulness measures from [8] and [16], 3.b Color related correlates from Color Appearance Models (CAM) [17, 18]. Colorfulness measures 3.a focus on low computational complexity:

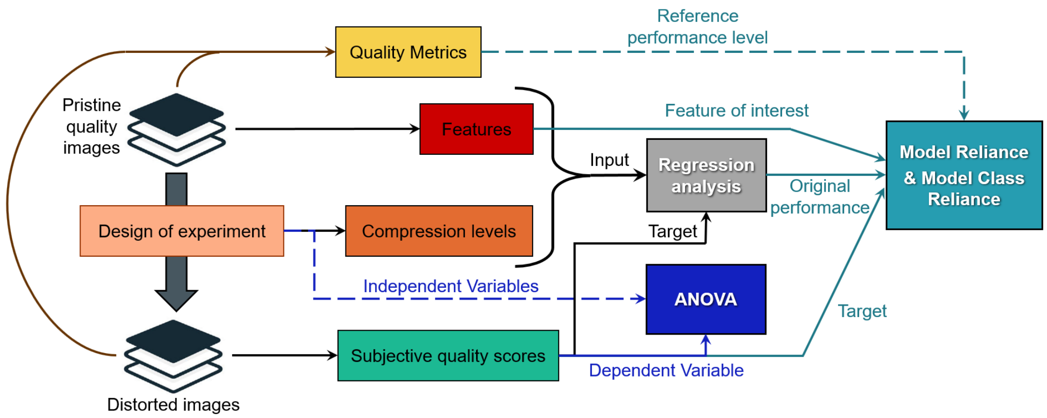


Figure 2: Pipeline of experiment. The analysis is done in 3 steps: ANOVA analysis on the subjective grades, regression analysis and calculation of the relevance of each feature through the Model Class Reliance approach.

they combine the mean and variance of the chroma planes of images in a simple color opponent colorspace[1] either linearly for M3 [8] or in the logarithmic domain to calculate the C1 and C2 measures in [16]. Color Appearance Models (CAM) model the early stages of human vision to predict the perception of images in specific viewing conditions. We include as features the colorfulness (M), chroma (C) and saturation (s) correlates from CIECAM02, CIECAM16, Hellwig22 and ZCAM [17, 18] [2].
The viewing conditions (ambient light + display) are modeled as best possible given the available information. For experiments where the precise information is not available (e.g. crowdsourcing) standard viewing parameters are assumed.

Regression analysis

For the regression analysis, the features presented in the previous section are calculated on the original quality images and then used together with the compression levels to build regression models predicting quality evaluations. We denote the regression models $f_k(X_1, X_2)$ where X_2 is the variable coding the Compression level, X_1 is one of the features presented in Sec. *Chosen features* to characterize the content and f_k the regression model. Two types of models, chosen for their simplicity and therefore robustness, are used: Ordinary Least Squares (OLS) [14] and K-Nearest Neighbors (KNN). OLS regression is rarely used because of its sensitivity to multicollinearity between features, so it is fitting here when there is no such risk as one input feature varies only by compression level whereas the second varies only by content. All modeling is done by splitting the dataset in 5 folds: at each iteration, training is done on 80% of the images and the remaining 20% are used for testing with no content present in both training and testing. The regression performance is evaluated through Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SROCC).

Estimation of the feature importance

The first step of the *Empirical Model Reliance* method is to define a ”model of interest” which performance will serve as reference to establish a threshold for ”well-performing models”. There is no consensual ”model of interest” for quality prediction of compressed image, and instead we use two reference models as indicators of performance: a full-reference (FR) quality metric and the OLS model built using only X_2 (Compression level) as input. We calculate several state-of-the-art FR quality metrics as they are currently the best performing type of models for image quality predictions and we use the metric performing best as indicator. Given that FR quality metrics and the f_k have access to different types of data, it is not possible to directly compare their respective performance. The OLS model using only X_2 (Compression level) as input represents the baseline of linear prediction without knowledge about the content that we are building on, it is in that sense the lower performance threshold. For each model f_k, we first calculate the *Empirical Model Reliance* $\widehat{MR}(X_1, f_k)$, i.e. how much this specific model (with fixed coefficients) relies on the value of the feature of interest X_1 for its performance. Our feature of interest X_1 is only related to the original content and does not depend on the compression level, therefore the expected loss when introducing permutations, Eq. 3.3 in [7], can be rewritten in our case as:

$$\hat{e}_{switchX_1}(f_k) = \frac{1}{n_{cpv}(n_{ct}-1)} \sum_{i=1}^{n_{cpv}} \sum_{ct_j \neq ct_i} \left(y_j - f_k(X_{1,i}, X_{2,j})\right)^2 \qquad (1)$$

where n_{cpv} is the number of compressed versions for each content, ct_i is the original content corresponding to stimuli i and $Y = \{y_j, j \in [\![1, n_{Samples}]\!]\}$ is the target variable. For the OLS model case, the regression model is defined by $\beta = (\beta_1, \beta_2)$, so following the same development as Eq. 3.3 to 7.2 in [7], Eq. 1 can be rewritten as

$$\begin{aligned}\hat{e}_{switchX_1}(f_k) = \frac{1}{n_{cpv}} \Bigg(& Y'Y - 2\begin{bmatrix} X_1'\mathbf{W_{bk}}Y \\ X_2'Y \end{bmatrix}\beta \\ & + \beta' \begin{bmatrix} X_1'X_1 & X_1'\mathbf{W_{bk}}X_2 \\ X_2'\mathbf{W_{bk}}X_1 & X_2'X_2 \end{bmatrix}\beta \Bigg)\end{aligned} \qquad (2)$$

where matrices are noted in bold font and capital letters are used for vectors. Eq. is similar to Eq. 7.2 in [7], with the replacement of the matrix $\mathbf{W}$, by the matrix $\mathbf{W_{bk}} = \frac{1}{n_{ct}-1}\mathbf{1_n}\mathbf{1_n'} - \mathbf{B_{bk}}$ where $\mathbf{B_{bk}} \in \mathbb{R}^{n_{cpv} x n_{cpv}}$ has for elements $B_{bk}(i,j) = \delta(ct_i, ct_j)$. The lower and upper bounds for $\widehat{MR}$, empirical Model Class Reliance $\widehat{MCR}_+$ and $\widehat{MCR}_-$, are calculated following the procedure for linear models from [7] with an adaptation of Eq 7.3 [7] to our

[1] (R-G, (R+G)/2 - B)

[2] Calculated with Colour https://colour.readthedocs.io/en/develop/

Table 1: Main characteristics of the datasets used

Name	# contents # samples	Colorspace	Codec	Subjective evaluations
Kadid10k [19]	81 / 810	sRGB	JPEG2000 / JPEG	DCR visible ref. - 30 eval. per PVS / Crowd-sourcing
CID 22 [20]	49 / 1512	sRGB	JPEG / JPEG2000 / JPEG-XL	Adapted PC - $\geq$49 eval. & DSIS - 101 eval. / Controlled & Crowdsourcing
TID-UPIQ [21, 5]	25 / 250	sRGB	JPEG2000 / JPEG	Pairwise Comparison - $\approx$ 38 eval. / Controlled & Crowdsourcing
Narwaria-UPIQ [22, 5]	10 / 140	HDR/ BT.709	TMO / JPEG / iTMO - TMO approx. of iCAM06 using MSE/SSIM	ACR-HR - 26 eval. per PVS, scaled to JOD / Controlled
Korshunov-UPIQ [1, 5]	20 / 240	HDR/ BT.709	JPEG-XT at 3 profiles	DSIS - 22 eval. per PVS, scaled to JOD / Controlled
IRISAWCG4K [6]	8 / 96	HDR/ WCG (BT.2020)	HEVC - 3 Chroma settings: 8b, 10b with and 10b W/O Chroma QP offset	DSIS visible ref. - 13 eval. per PVS/ Controlled

case following Eq. 1. For linear models as used in this study, relaxing the fit to bound the range of $\widehat{MR}$ values means allowing variation in the β coefficient vector that define the model. $\widehat{MCR}_{+}$ and $\widehat{MCR}_{-}$ are determined through quadratic expression on β (and solved via a quadratic solver).

Experimental results

Dataset

The method was tested on six publicly available datasets of subjective evaluations of compressed images. The main characteristics of the datasets used in the experiment are summarized in Table 1. For the datasets containing other types of defects than compression-based ones, such as Kadid10k [19], CID22 [20] and TID2013 [21], we only selected the subset of distortions induced by standardized image compression: JPEG, JPEG2k and JPEG-XL. Three datasets [22, 1, 6] are focused on compression of HDR content with BT.709 primaries for the first two and WCG for the last one. In [22] and [1], the authors process HDR images through tone mapping, JPEG compression and inverse tone mapping. In the first study Narwaria et al. optimized iCAM06 TMO/iTMO via MSE or SSIM and retain only the SDR version in the compression part. In [1], Korshunov et al. use the three profiles of the JPEG-XT standard to create an extension layer and reconstruct the HDR part after the compression of both base and extension layers via JPEG. For the regression analysis, the psychometric scaling of subjective grades to align them on a similar scale performed in [5] is used. Finally, in [6], Rousselot et al. apply HEVC compression on images in WCG (BT.2020). Their focus is specifically on the importance of color and the 3 compression settings they use differ by the handling of the chrominance channels.

ANOVA analysis for respective effect size of content vs compression level

N-way ANOVA analysis was performed to evaluate the significance of the different independent variables as factors, i.e. calculate the % of variance from the subjective grades due to *content* and *compression level* respectively. This is done on the three HDR datasets as the grades per observer are available: Narwaria, Korshunov and IRISAWCG4K. N-way ANOVA is applied on the following independent variables: Compression level (CpLvl), content (Cnt) and compression type (CpTyp). The analysis is first calculated with all factors and their interactions, and then a second time with keeping only the factors and interactions which have a statistically significant effect on the measured variable. Results are detailed in Table 2. For all datasets, the factors CpLvl and Cnt have a significant influence on the subjective grades as factors and through their interaction. It is also to be noted that for all datasets, the type of compression settings has little or no significant impact on the grades (see column Codec in Tab. 1 for the list of specific settings for each dataset). The desired outcome of this analysis is the comparison of the effect size, reported as ω^2 values in Table 2 [14]. The proportion of variance explained by the full model is reported in the right-most column: it ranges from 65% for *IRISAWCG4K* to 78% for *UPIQKorshunov*. In terms of model performance, this global ω^2 is equivalent to the adjusted coefficient of fit R2, meaning that the corresponding PCCs are 0.824, 0.887 and 0.801 respectively for *UPIQNarwaria*, *UPIQKorshunov* and *IRISAWCG4K*. The proportion of variance form the subjective grades explained by CpLvl / Cnt are 45.5% / 10.1%, 62.4% / 7% and 49% / 6.9% respectively for *UPIQNarwaria*, *UPIQKorshunov* and *IRISAWCG4K*.

Estimating feature importance via regression models

Performance analysis - The MCR method necessitates a reference level for the performance of the considered regression models. We have calculated two models that differ from the feature + CpLvl models: the OLS model built with CpLvl only and state-of-the-art image quality metrics. The selected metrics are: PSNR-Y, PSNR-RGB, SSIM, MS-SSIM (with PU coding as pre-processing for HDR content [5]) and the more complex ColorVideoVDP [23]. For all datasets, the best performing quality metric is ColorVideoVDP (CVVDP), and it is the only one reported here for space reasons. ColorVideoVDP and OLS models use very different information: both original and degraded content for the full reference metric vs. original content and compression levels for the OLS-R models. Therefore a direct comparison is not relevant but the image quality metric is more used here as a reference of performance level. The performance in terms of SROCC for each OLS model is visible in Fig. 3 for each dataset separately. The PCC and SROCC performances of the best QM, the OLS with only CpLvl as input variable and the OLS and KNN model with 1 content feature and CpLvl with highest performance are given in Table 3. Modeling with KNN or OLS regression yields similar performance levels. In terms of SROCC, the OLS-R models with 1 content feature and CpLvl achieve better performance than the

Table 2: ANOVA analysis of the independent variables compression level and content

Dataset name	Compression level - CpLvl df / F / p	ω^2	Content - Cnt df / F / p	ω^2	Int. CpLvl x Cnt df / F / p	ω^2	Model w. all significant factors Total ω^2	Factors w. $\omega^2 \geq 1\%$
UPIQNarwaria	6 / 829.5 p=0	0.455	9 / 124.0 p=6.4E^{-202}	0.101	54 / 7.5 p=7.1E^{-52}	0.032	0.679	CpTyp x Cnt (0.07) CpTyp (0.010)
UPIQKorshunov	3 / 5148 p=0	0.624	19 / 94.3 p=0	0.07	57 / 15.9 p=2E^{-139}	0.034	0.787	CpTyp x Cnt (0.03) CpLvl x CpTyp x Cnt (0.014)
IRISAWCG4K	3 / 584.5 p=8.2E^{-231}	0.49	7 / 36.4 p=3.53E^{-46}	0.069	21 / 2.2 p=0.001	0.007	0.651	CpTyp x Cnt (0.073)

OLS-R model using only CpLvl for every dataset but this difference is statistically significant only for the UPIQTID dataset.

Model Reliance - Each of the models considered has two input features: the compression level and a feature calculated on the original content that is our *feature of interest* in the sense of [7]. The Empirical Model Reliance $\widehat{EMR}$ values are presented for each OLS model in Fig 3 in the lower part of the y-axis, separately for each dataset. The SROCC of the OLS model indicates which features are the most useful for predicting the part of the subjective grades not due to the compression level. Interesting prediction models are those achieving both better performance than CpLvl alone and better other features. The $\widehat{EMR}$ measures how much a model relies on the feature of interest for the prediction when being best fitted to the current dataset, and therefore the SROCC performance value and the $\widehat{EMR}$ should be read jointly. If we consider for instance the features C2 and CIE02_s for the dataset TID2013, the corresponding OLS models achieve similar SROCC performances of 0.869 (ranking as 6th highest SROCC among 21) and 0.863 (ranking as 10th highest SROCC) respectively. However, their $\widehat{MR}$ values are 1.008 for C2 and 1.098 for CIE02_s. Those values indicate that the OLS-C2 model actually does not rely much on the feature C2 as the MSE increases by less than 1% when introducing noise in the said feature. On the contrary the prediction error of the OLS-CIE02_s increases by almost 10% when the values of CIE02_s are permuted.

Model Class Reliance - The results for Empirical Model Class Reliance are presented in Fig. 4 where the feature importance is represented on the x-axis in terms of model reliance values and the model performance is represented on the y-axis in terms of MSE loss. The $\widehat{MR}$ is represented by the $+$ symbols, it indicates the reliance value for the model best fitted for a dataset. For the $\widehat{MR}$, the most desirable points are those closer to the bottom right corner of the plots (lower loss and higher reliability). The upper and lower bounds of MCR, $\widehat{MCR}_+$ and $\widehat{MCR}_-$, are the left and right parts of the curves respectively starting from that $+$ symbol. The $\widehat{MCR}_+$ and $\widehat{MCR}_-$ indicate the range within which the empirical model reliance $\widehat{MR}$ could evolve when relaxing the constraint of minimal loss, i.e. relaxing the constraint of best fit to the subjective data. As expected in every case the $\widehat{MCR}_+$ and $\widehat{MCR}_-$ curves grow further apart when the minimal loss (y-axis) increases. The threshold below which a model does not rely on a variable is the black line at EMR=1 and the threshold above which a model does not improve on the OLS model with CpLvl only is the horizontal dashed line. The most interesting features are those whose $\widehat{MCR}_-$-$\widehat{MCR}_+$ curves are the closest to the bottom right part of the plots and with the "flattest" shape for the upper bound $\widehat{MCR}_+$, i.e. indicating a higher robustness with regard to the fit to the specific testing data.

the The plots 4a, 4b, present results for features from Group 1 and DR from Group 2, and the plots 4c, 4d, for a selection of features from Group 3 (a and b): $M3H03$,$CIE16_s$, $ZCAM_s$, $Hellwig_M$ and $Hellwig_s$. For UPIQTID, features from group 1 are the most useful and SI performs comparatively to the block contrast tried. For the UPIQNarwaria dataset, the DR feature is among the most interesting which is sensible given the content is HDR as well as color-related measures M3 and saturation from CAMs CIE02_s, CIE16_s and ZCAM_s.

Conclusion and future work

This paper investigates the role of original content in visual perception of compressed images, in contrast to that of the compression level. By using ANOVA analysis on existing datasets, we show that the respective proportion of variance in the subjective evaluations is comprised between 45-62% for the compression level and 7-10% for the image content. Secondly, we present a framework building on regression analysis to robustly determine which features characterize well image content with the Model Class Reliance approach . Two measures are added to the traditional ranking of features based on the performance of the corresponding regression models: the Empirical Model Reliance $\widehat{MR}$ estimates how much a regression model relies on the considered feature when fully optimizing for a training dataset and the Empirical Model Class Reliance provides upper and lower bounds for $\widehat{MR}$ when relaxing the loss optimization. Comparing results for the SDR and HDR datasets studied shows that the most useful features for SDR are SI/block contrast measures whereas other aspects such as DR and color features are most relevant for HDR content. An extension of the work to evaluate whether across viewing conditions influences the results, as well as to refine the measure of robustness depending on the compression level is planned.

References

[1] P. Korshunov, "Subjective quality assessment database of HDR images compressed with JPEG XT," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2015.

[2] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2023.

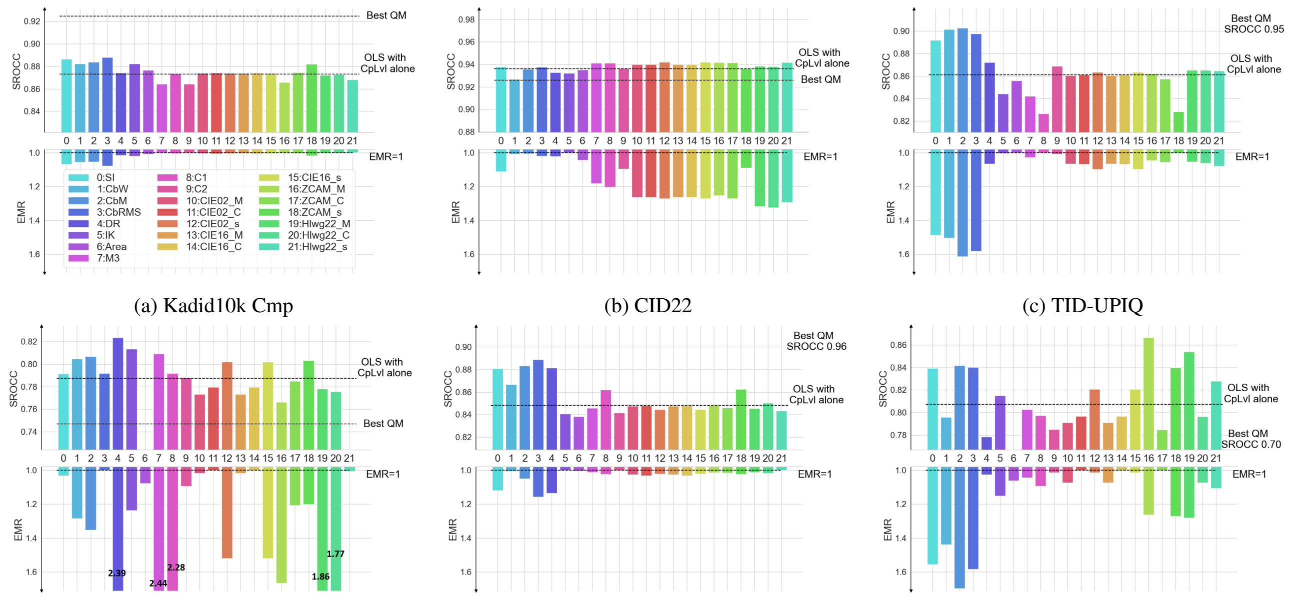
[3] S. Winkler, "Analysis of Public Image and Video Databases for Quality Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, 2012.

[4] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?," in *Fourth International Workshop on Quality of Multimedia Experience*, IEEE, 2012.

[5] A. Mikhailiuk, M. Perez-Ortiz, D. Yue, W. Suen, and R. Mantiuk, "Consolidated Dataset and Metrics for High-Dynamic-Range Image

Table 3: Prediction of quality for each dataset through different methods: best image quality metric, OLS regression using only CpLvl and regression models (OLS and KNN) build using 1 feature calculated on original content and the compression levels.

Dataset name		**Kadid10kCmp**	**CID22**	**UPIQTID**	**UPIQNarwaria**	**UPIQ Korshunov**	**IRISAWCG4K**
Best QM	Metric	CVVDP	CVVDP	CVVDP	CVVDP	CVVDP	CVVDP
	PCC / SROCC	0.938 / 0.924	0.856 / 0.926	0.956 / 0.946	0.774 / 0.747	0.934 / 0.962	0.681 / 0.73
OLS CpLvl	PCC / SROCC	0.903 / 0.873	0.937 / 0.936	0.842 / 0.861	0.799 / 0.787	0.877 / 0.848	0.838 / 0.807
Best performing models using CpLvl + 1 feature on original content							
OLS	Feature	CbRMS	CIE16_s	CbM	DR	SI	CbM
	PCC / SROCC	0.903 / 0.883	0.94 / 0.942	0.874 / 0.903	0.853 / 0.859	0.871 / 0.879	0.853 / 0.854
KNN	Feature	CbRMS	CIE02_C	CbRMS	DR	CbM	CbW
	PCC / SROCC	0.946 / 0.875	0.941 / 0.941	0.918 / 0.895	0.823 / 0.825	0.882 / 0.87	0.849 / 0.845



(a) Kadid10k Cmp (b) CID22 (c) TID-UPIQ

(d) Narwaria-UPIQ (e) Korshunov-UPIQ (f) IRISAWCG4K

Figure 3: Performance of OLS regression for each feature given as SROCC on the upper part of the y-axis and Empirical Model Reliance ($\widehat{EMR}$) on the lower part of the y-axis (the axis is reversed, longer bars means higher values). On the upper half, the performance of the best performing quality metric (ColorVideoVDP) and the model built using only Cp Lvl are also given for comparison. On the lower half, the line of $\widehat{EMR}$=1 indicating the threshold under which the model does not rely on the feature is given for reference.

Quality," *IEEE Transactions on Multimedia*, 2022.

[6] M. Rousselot, "Quality Assessment of HDR/WCG Images Using HDR Uniform Color Spaces," *Journal of Imaging*, vol. 5, 2019.

[7] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, 2019.

[8] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human Vision and Electronic Imaging VIII*, 2003.

[9] V. Hulusic, K. Debattista, G. Valenzise, and F. Dufaux, "A model of perceived dynamic range for HDR images," *Signal Processing: Image Communication*, vol. 51, 2017.

[10] M. Narwaria, "An objective method for High Dynamic Range source content selection," in *2014 6th International Workshop on Quality of Multimedia Experience, QoMEX 2014*, 2014.

[11] P. Hanhart and T. Ebrahimi, "Calculation of average coding efficiency based on subjective quality scores," *Journal of Visual Communication and Image Representation*, vol. 25, 2014.

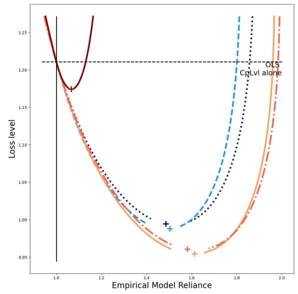
[12] H. Yu and S. Winkler, "Image complexity and spatial information," in *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2013.

[13] A. Choudhury and S. Daly, "HDR Display Quality Evaluation by incorporating Perceptual Component Models into a Machine Learning framework," *Signal Processing: Image Communication*, 2019.

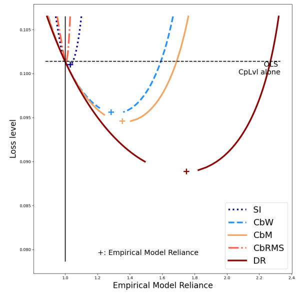
[14] Cohen J., Cohen P., West S., and Aiken L., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge, 2013.

[15] L. Krasula, K. Fliegel, and P. Le Callet, "FFTMI: Features Fusion for Natural Tone-Mapped Images Quality Evaluation," *IEEE Transactions on Multimedia*, vol. 22, 2020.

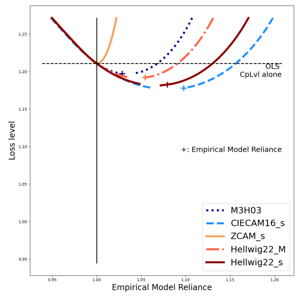
[16] K. Panetta, C. Gao, and S. Agaian, "No reference color image contrast and quality measures," *IEEE Transactions on Consumer Elec-



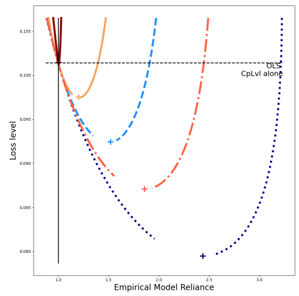
(a) Feature set 1 - UPIQTID.



(b) Feature set 1 - UPIQNarwaria.



(c) Feature set 2 - UPIQTID.



(d) Feature set 2 - UPIQNarwaria.

Figure 4: $\widehat{MR}$ and bounds $\widehat{MCR}_+$ and $\widehat{MCR}_-$ when relaxing the loss at different values for two set of features and datasets UPIQTID and UPIQNarwaria). Feature set 1 is composed of: *SI*, *CbW*, *CbM*, *CbRMS* and *DR*, Feature set 2 is composed of: $M3H03$,$CIE16_s$, $ZCAM_s$, $Hellwig_M$ and $Hellwig_s$. The vertical line at $\widehat{MR}$=1 indicates the threshold below which features have no real impact on the model and the horizontal dashed line indicates the loss when modeling with CpLvl only.

tronics, vol. 59, 2013.

[17] L. Hellwig and M. D. Fairchild, "Brightness, lightness, colorfulness, and chroma in CIECAM02 and CAM16," *Color Research & Application*, vol. 47, 2022.

[18] M. Safdar, J. Y. Hardeberg, and M. Ronnier Luo, "ZCAM, a colour appearance model based on a high dynamic range uniform colour space," *Optics Express*, vol. 29, 2021.

[19] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A Large-scale Artificially Distorted IQA Database," in *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2019.

[20] J. Sneyers, E. Ben Baruch, and Y. Vaxman, "AIC-3 Contribution from Cloudinary: CID22," ISO/IEC JTC 1/SC29/WG1, 2023.

[21] N. Ponomarenko, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, 2015.

[22] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, 2013.

[23] R. K. Mantiuk, P. Hanji, M. Ashraf, Y. Asano, and A. Chapiro, "ColorVideoVDP: A visual difference predictor for image, video and display distortions," *ACM Transactions on Graphics*, vol. 43, 2024.